Weekly Report
08/17/2015 - 08/23/2015

Jing Xia

August 24, 2015

1 Summary

This week I mainly focus on the DataScanner project.

2 Projects

2.1 Project 1 - DataScanner1

This week I implemented a little more features to enhance the system. The view selector widget now are able to select/de-select 1D views and 2D views. Thus the tree can only show 1D or 2D views, which makes it easier to do dimension reduction.
I'm also working with Zongzhuang Li and Sidong Wang on dendrogram layout of the quartet view tree. Currently views are still overlapping with each other. We worked out a simple depth-based force layout to address the overlapping problem. Result to be expected next week.
According to review comments, I've started paper revising and done Abstract and Introduction Section. The major change that I avoid the dissection concept and focus on the technique.

Abstract

Exploring multi-dimensional datasets can be cumbersome if data analysts have little knowledge about the data. Various programming tools and visualization tools have been invented for efficient data examining and understanding. However, the needed workload varies largely with respect to data complexity and user expertise, which can only be reduced with rich background knowledge over the data. In this paper we address the workload challenge with the DataScanner method that affords information detection over multi-dimensional data and that serves as the background knowledge support for further in-depth algorithms or visual designs. We contribute a novel data organizing scheme that leverages an information-theoretic view matching algorithm to uncover information- aware relations among different data views, and thereby discloses redundant or highly

related dimensions. We propose an expressive view exploration technique to allow for adaptive and interest-driven investigation of data views. The integrated system, DataScanner, empowers analysts with rich user controls to interactively detect the view-wise structure of multi-dimensional data.

Introduction

Although multi-dimensional data is considered of great value, without proper processing and interpretation it is only garbage or even poison (considering misleading conclusions generated from data). In particular, multi-dimensional data exploration shares a common need of find- ing relations or frequent patterns between dimensions and a custom need of semantic exploration. However, data analysts are often uninformed before they make all attempts to understand the data. Typical challenges for multi-dimensional data exploration are:

- Data analysts often feel confused when handling a large-sized multi-dimensional dataset. They explore the dataset by randomly navigating from one view to another, expecting to find useful pat- terns. They want to make the best use of visualization, but an overview of all data views without proper organization is overwhelming.

- Redundant data takes unnecessary analysis workload. Even though the data is cleaned in terms of data format, there might be redundant dimensions that encode useful but duplicate information. If analysts can notice the similarity or the relationships across the dimensions, the cognitive complexity can be reduced.

Programming languages or data processing tools like Python, R or SQL are well adopted in multi-dimensional data exploration. How- ever, programming tools suffers from two drawbacks. First, it requires programming skills; second, it lacks a way to draw support from data analysts? expertise due to poor support of expressive data feedback. Alternatively, visualization layouts like parallel coordinates and scatter plot matrix can be employed to depict dimension-wise data distributions expressively. Though effective, its performance is heavily dependent on the data complexity and the user expertise. Both approaches can generate a set of data views for deep investigation, but without an appropriate organization, analysts have to randomly explore the many views in pursuit of valuable information.

These challenges can probably be leveraged with an informative data context that depicts the basic distributions of data dimensions as well as the relations between them. This elicits the demand for an efficient information-aware organization and exploration scheme, whose primary task is to provide an information-rich understanding of the data and to give clues of redundant dimensions. And we meet the demand by backing up visualization with not only data processing but also data organizing techniques. The data organizing step tries to re- veal basic structures between data dimensions and to guide analysts? exploration process.

In this paper we contribute the design and implementation of a data organizing and exploration scheme that upgrades conventional multi-dimensional data

exploration process with a novel information- theoretic view matching and organizing technique. We organize extracted views as a configurable tree structure by means of a quartet analysis technique [31], presenting the relations among data views.
The main reason for using a quartet-based tree is that quartets are particularly efficient in cases that a global view may not be immediately deduced from the raw data, but on smaller scales, local relations may be discerned more reliably. In addition, the constructed categorization tree results in an information-aware tree structure of data views, where views with high relations are near to each other and views with low relations are far apart. The implemented system, DataScanner, integrates view matching and organizing techniques and provides an expressive visualization for efficient navigation. It empowers analysts with comprehensive preview and rich controls over various views. In summary, the contributions of this paper are as follows:

- A view matching and organizing scheme that generate structures of high-dimensional data;
- An exploration process that supports interactive investigation and view-wise relation validation of high-dimensional data.

3 Paper Reading and Miscellaneous

I read the pitfall paper by Tamara.

4 To Do List

1. Go on with DataScanner system implementation
2. Go on with paper writing